Question answering in natural language using common sense

IBM Watson winning Jeopardy in 2011

Example questions from Jeopardy with Watson

- It's just a bloody nose! You don't have this hereditary disorder once endemic to European royalty : *Haemophilia*
- You just need a nap! You don't have this sleep disorder that can make sufferers nod off while standing up : *Narcolepsy*
- Heitor Villa-Lobos dedicated his "12 Etudes" for this instrument to Andres Segovia *: Guitar*
- Paganini's "24 Capricci" set the standard for etudes for this instrument *: Violin*
- Rembrandt's Biblical Scene "Storm on the Sea of" this was stolen from a Boston museum in 1990 : *Galilee*

GPT2, the text-completion system of OpenAI

System Prompt (human-written)

- In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

Model Completion (machine-written, 10 tries)

- The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science. Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

O. Etzioni, CEO of the Allen Institute for AI:

No AI system currently deployed can reliably answer a broad range of simple questions such as:

- If I put my socks in a drawer, will they still be in there tomorrow?
- How can you tell a milk carton is full?

Authors of one of the top question answering systems say that

... the following is still not satisfactorily answered:

„The trophy would not fit in the brown suitcase because it was too big. What was too big?“

Common sense

Wikipedia, following Ernest & Marcus: "Commonsense reasoning is one of the branches of artificial intelligence (AI) that is concerned with simulating the human ability to make presumptions about the type and essence of ordinary situations they encounter every day.

These assumptions include judgments about the physical properties, purpose, intentions and behavior of people and objects, as well as possible outcomes of their actions and interactions."

Commonsense knowledge

Wikipedia: "In artificial intelligence research, commonsense knowledge consists of facts about the everyday world, such as 'Lemons are sour', that all humans are expected to know.

It is currently an unsolved problem in Artificial General Intelligence and is a focus of the Paul Allen Institute for Artificial Intelligence."

The current approaches to CSR

Broadly classified as based on either

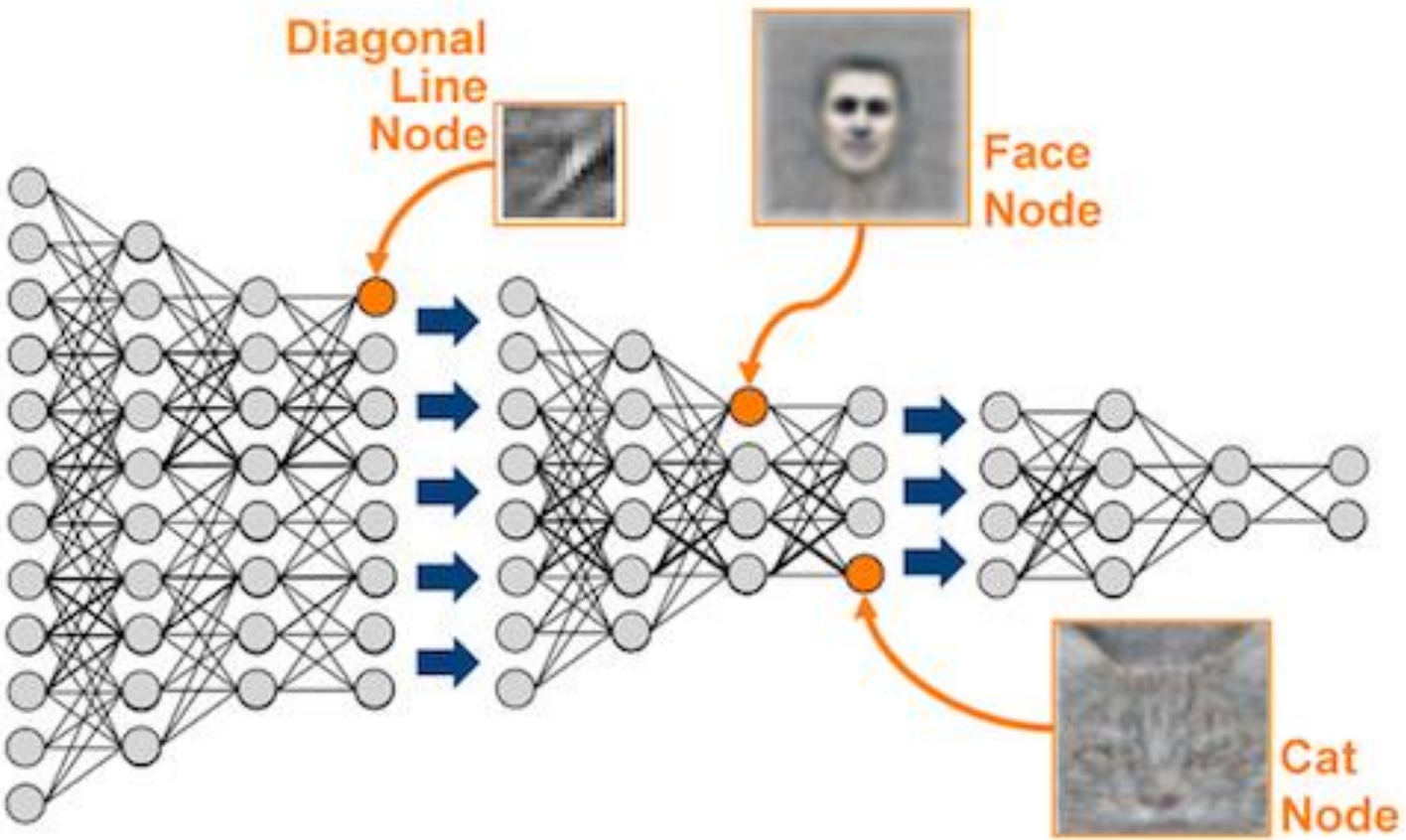
- machine learning (ML) on a large corpora of natural language texts or
- logical reasoning

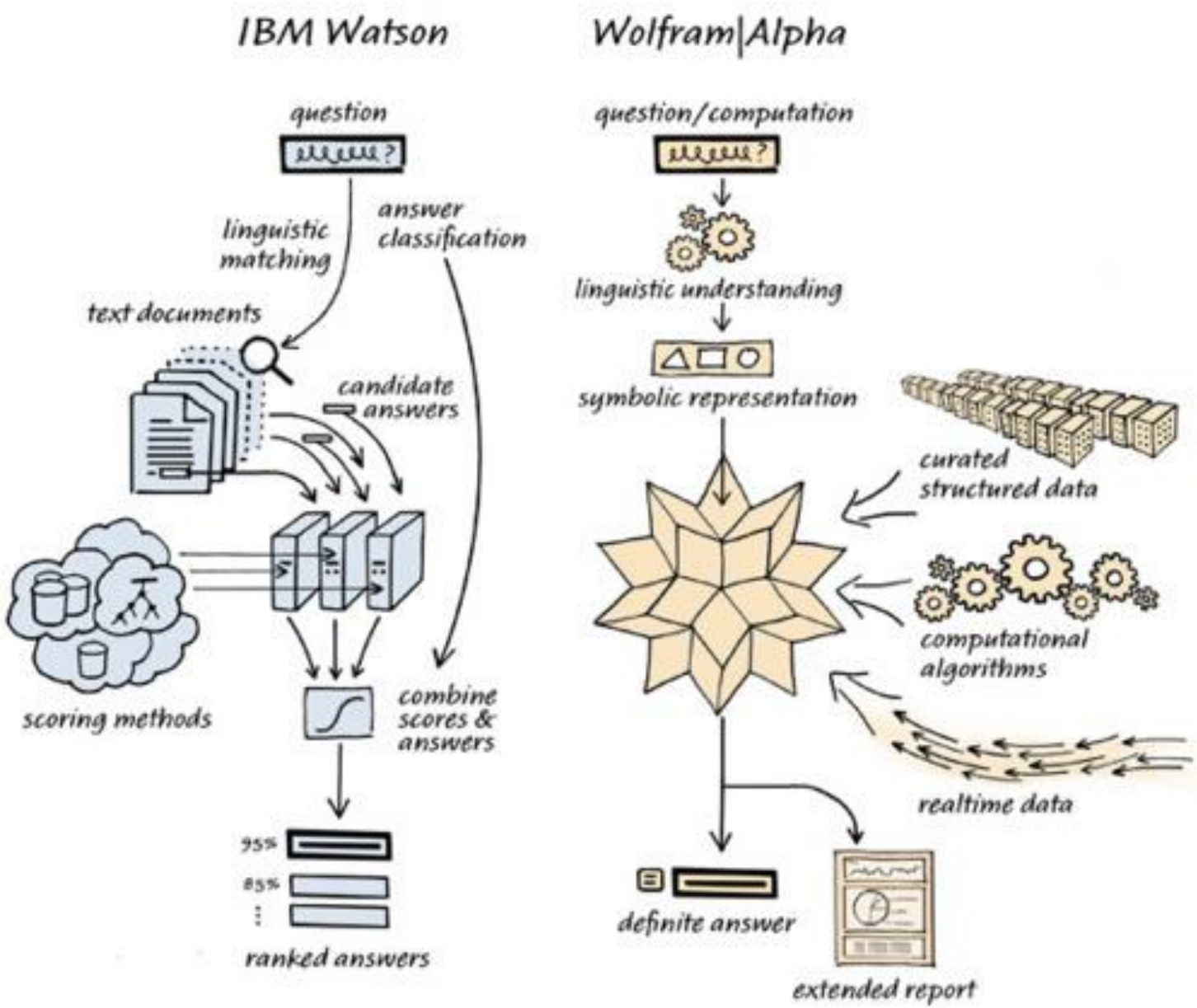
Main successes achieved by ML systems

The core method for ML approaches is to build a prediction system using ML techniques, which, given an input text, predicts the next or missing word or phrase.

The most prominent current underlying system is BERT from Google AI, with a number of specialised systems built on BERT.

Neural networks




IBM Watson
question
linguistic matching
answer classification
text documents
candidate answers
scoring methods
combine scores & answers
95%
85%
ranked answers
Wolfram|Alpha
question/computation
linguistic understanding
symbolic representation
curated structured data
computational algorithms
realtime data
definite answer
extended report

Question answering benchmarks

Winograd schema

The city councilmen refused the demonstrators a permit because they [feared/advocated] violence.

SuperGlue benchmark

An example with a binary choice question:

- Premise: My body cast a shadow over the grass.
- Question: What is the CAUSE for this?
- Alternative 1: The sun was rising.
- Alternative 2: The grass was cut..

For these kinds of questions the authors note that the best of their systems, BERT++, achieved ca 74% accuracy, while people achieved 100%.

ARC benchmark

Which property of a mineral can be determined just by looking at it?

(A) luster [correct]

(B) mass

(C) weight

(D) hardness.

The current leader of the ARC public leaderboard is the FreeLB-RoBERTa (single model) system, with the accuracy 68% and several BERT-based systems following closely.

OpenBookQA

Which of these would let the most heat travel through?

- a new pair of jeans.
- a steel spoon in a cafeteria.
- a cotton candy at a store.
- a calvin klein cotton hat.

Answering the question requires using the given science fact ``Metal is a thermal conductor''and common knowledge expected to be present or derived from the knowledge base of the measured system: ``Steel is made of metal. Heat travels through a thermal conductor.''

The current public leaderboard for OpenBookQA has the system AristoRoBERTaV7 of the authors achieving best performance: 78% accuracy.

Some G. Marcus new benchmark examples

None of the current top-of the line systems can give a satisfctory answer to these:

- There are six frogs on the log. Two leave, but three join. The number of frogs on the log is now?
- Yesterday I dropped my clothes at the dry cleaners and have yet to pick them off. Where are my clothes now?

Marcus summarizes the findings thus: ``large-scale language models do a good job of figuring the topic under consideration, and what the plausible set of masked words / continuations might be, given the input context. But a poor job of reasoning about which specific response is the right one.''

Logical approaches

CYC project: 1985 - …. ongoing

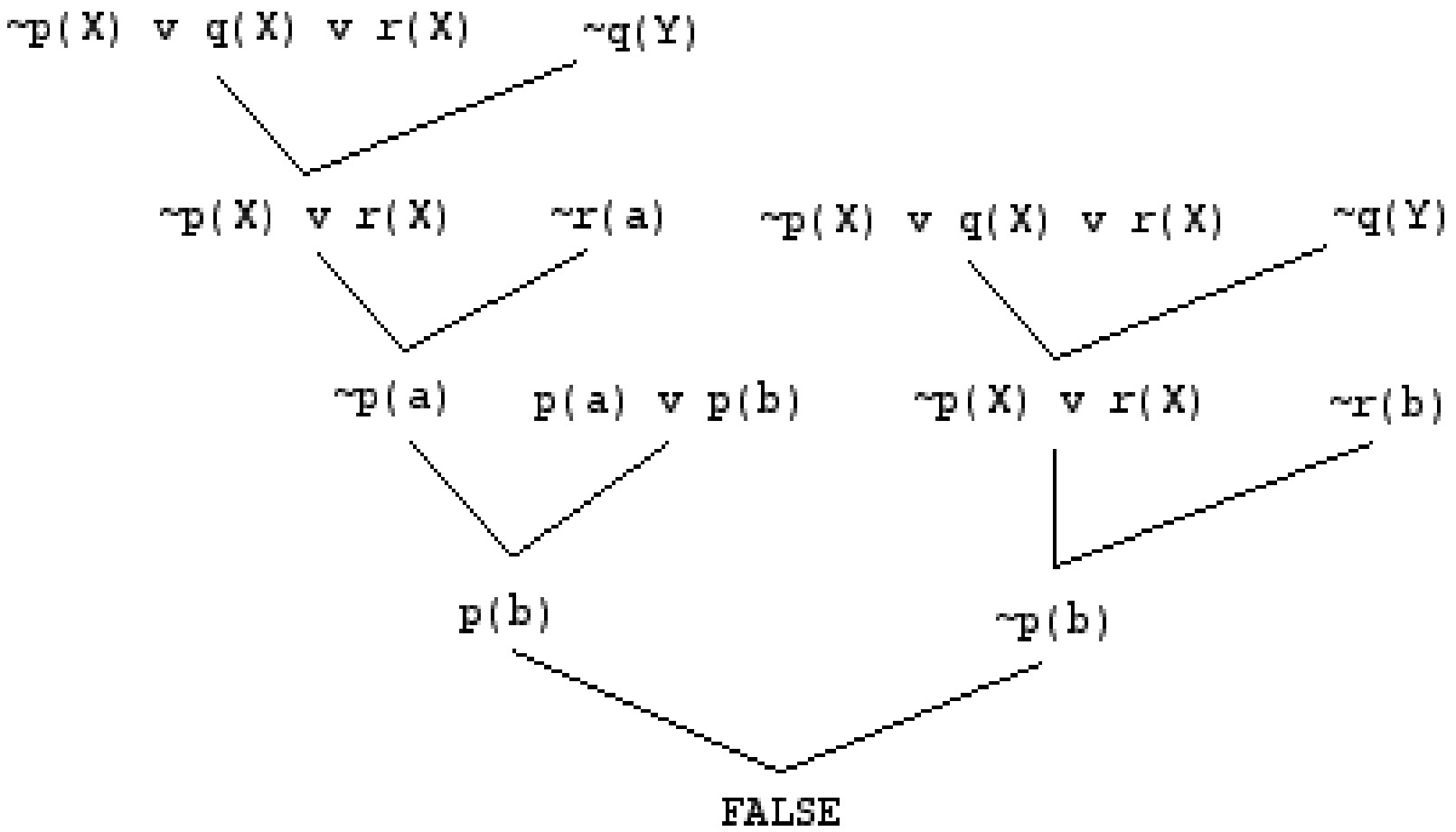
Developed the largest known manually made commonsense knowledge base along with reasoners specially built for this knowledge base.

Parts of the knowledge base -- OpenCyc -- are available as a large FOL axiom set.

CYC deduction problems in remain relatively hard for top-of-the-line automatedreasoners, mostly due to little focus on CYC-specific proof search heuristics.

Despite several successes, the approach taken in the CYC project has been often viewed as problematic) and has been repeatedly used as an argument against logic-based methods in CSR.

Deduction tree



Robbins algebras are boolean: Mccune, 1997

In 1933, E. V. Huntington presented the following basis for Boolean algebra:

$x + y = y + x.$

$(x + y) + z = x + (y + z).$

$n(n(x) + y) + n(n(x) + n(y)) = x.$

Shortly thereafter, Herbert Robbins conjectured that the Huntington equation can be replaced with a simpler : $n(n(x + y) + n(x + n(y))) = x.$ Robbins and Huntington could not find a proof, and the problem was later studied by Tarski and his students.

The successful search took about 8 days on an RS/6000 processor and used about 30 megabytes of memory.

2 (wt=7) [] -(n(x + y) = n(x)).
3 (wt=13) [] n(n(n(x) + y) + n(x + y)) = y.
5 (wt=18) [para(3,3)] n(n(n(x + y) + n(x) + y) + y) = n(x + y).
6 (wt=19) [para(3,3)] n(n(n(n(x) + y) + x + y) + y) = n(n(x) + y).
24 (wt=21) [para(6,3)] n(n(n(n(x) + y) + x + y + y) + n(n(x) + y)) = y.
47 (wt=29) [para(24,3)] n(n(n(n(n(x) + y) + x + y + y) + n(n(x) + y) + z) + n(y + z)) = z.
48 (wt=27) [para(24,3)] n(n(n(n(x) + y) + n(n(x) + y) + x + y + y) + y) = n(n(x) + y).
146 (wt=29) [para(48,3)] n(n(n(n(x) + y) + n(n(x) + y) + x + y + y + y) + n(n(x) + y)) = y.
250 (wt=34) [para(47,3)] n(n(n(n(n(x) + y) + x + y + y) + n(n(x) + y) + n(y + z) + z) + z) = n(y + z).
996 (wt=42) [para(250,3)] n(n(n(n(n(n(x) + y) + x + y + y) + n(n(x) + y) + n(y + z) + z) + z + u) + n(n(y + z) + u)) = u.
16379 (wt=21) [para(5,996),demod([3])] n(n(n(n(x) + x) + x + x + x) + x) = n(n(x) + x).
16387 (wt=29) [para(16379,3)] n(n(n(n(n(x) + x) + x + x + x) + x + y) + n(n(n(x) + x) + y)) = y.
16388 (wt=23) [para(16379,3)] n(n(n(n(x) + x) + x + x + x + x) + n(n(x) + x)) = x.
16393 (wt=29) [para(16388,3)] n(n(n(n(x) + x) + n(n(x) + x) + x + x + x + x) + x) = n(n(x) + x).
16426 (wt=37) [para(16393,3)] n(n(n(n(n(x) + x) + n(n(x) + x) + x + x + x + x) + x + y) + n(n(n(x) + x) + y)) = y.
17547 (wt=60) [para(146,16387)] n(n(n(n(n(x) + x) + n(n(x) + x) + x + x + x + x) + n(n(n(x) + x) + x + x + x) + x) + x) = n(n(n(x) + x) + n(n(x) + x) + x + x + x + x).
17666 (wt=33) [para(24,16426),demod([17547])] n(n(n(x) + x) + n(n(x) + x) + x + x + x + x) = n(n(n(x) + x) + x + x + x).

What we certainly need for practical commonsense reasoning

Efficiently query huge rule/fact databases
Accepting contradictory knowledge bases
Relevance measures and heuristics
Performing confidence calculations
Using default rules in a sensible way
Using analogues in reasoning

Topic-based confidence and relevance

Hobbits exist

{„Tolkien“: 100%, „fairytale“: 50%, „default“: 1%}

Possible villain?

Wolf {„fairytale“: 10%, „Tolkien“: 20%, „default“: 1%}

Saruman {„Tolkien“: 50%, „default“: 0%}

Classical default rule example

Birds can fly.

Penguins are birds.

Penguins cannot fly.

Pengu is a penguin.

Can Pengu fly?

A few more layers

Physical objects cannot fly.
Birds are physical objects.
Birds can fly.
Penguins are birds.
Penguins cannot fly.
Pengu is a penguin.
Can Pengu fly?

Nixon triangle

Nixon is a quaker and a republican.

Quakers are pacifists.

Republicans are not pacifists.

Is Nixon a pacifist?

Analogues

Kings are rich.

Kings are male.

Queens are like kings, but female.

Are queens rich?

GROCK framework under development

Instead of devising new specialized logics we propose a framework of extensions to the mainstream resolution-based search methods to make these capable of performing search tasks for practical commonsense reasoning with reasonable efficiency.

GK reasoner under development

The proposed extensions mostly rely on operating on ordinary proof trees and are devised to handle commonsense knowledge bases containing inconsistencies, default rules, taxonomies, topics, relevance, confidence and similarity measures.

We claim that machine learning is best suited for the the construction of commonsense knowledge bases while the extended logic-based methods would be well-suited for actually answering queries from these knowledge bases.

Moonshot goal:

build an efficient hybrid
ML + logic-based
NLP question answering system,
extend existing knowledge bases suitably,
increase reasoning efficiency by automatic
learning a la AlphaZero